Change the Product, Keep the Parameters

Associative Algebra Layers for Transformers

Paper authors: Ilya Koziev and Ivan Oseledets

Independent paper-reading presentation

arXiv:2609.32814v1 · 19 October 2026

A cheaper interaction rule retains the weight bank,
with a measured quality–throughput trade-off in a small LM.

The multiplication rule is an architecture choice

Fast matrix multiplication	This paper
Keep the target XW fixed.	Learn with $Y = \mu_{\mathcal{A}}(X, W)$.
Find a cheaper exact algorithm.	Choose a cheaper interaction table.
Example: Strassen uses 7 products for a 2×2 block product.	Keep all learned weight slots, while changing which products occur.

Question: how much computation can be removed before quality suffers?

Paper, Secs. 1–2, PDF pp. 1–3.

Two groups retain four weights with six products

$$X = \begin{bmatrix} A & U \\ V & D \end{bmatrix}, \qquad W = \begin{bmatrix} W_a & W_u \\ W_v & W_d \end{bmatrix}$$

$$X \star_2 W = \begin{bmatrix} AW_a & AW_u + UW_d \\ VW_a + DW_v & DW_d \end{bmatrix}$$

- Ordinary multiplication also has UW_v and VW_u on the diagonal.
- The new table removes these two interactions; every weight block remains.
- One token evaluates one row type: 3 block products instead of 4.

Paper, Eqs. (1)–(3), PDF pp. 1–2. Blocks denote compatible ordinary GEMMs.

The graph product attains its bilinear-rank bound

A vertex is a diagonal slot; an edge is an off-diagonal slot.

$$e_i e_i = e_i, \qquad e_i a_{ij} = a_{ij}, \qquad a_{ij} e_j = a_{ij}.$$

All other basis products vanish, including every edge–edge product.

Upper and lower bounds meet

$$R(\mathcal{Q}_G) \geq 2 \dim \mathcal{Q}_G - t(\mathcal{Q}_G) = 2(|V| + |E|) - |V| = |V| + 2|E|.$$

The explicit rule uses exactly $|V| + 2|E|$ products.

For the complete directed graph:

$$\dim \mathcal{Q}_q = q^2, \qquad \boxed{R(\mathcal{Q}_q) = 2q^2 - q}.$$

Paper, Secs. 2.2–3, Theorem 3.2, PDF pp. 3–4; proof in App. H.1, pp. 21–22. R: scalar bilinear rank; t: number of maximal two-sided ideals.

A token selects a row of the interaction table

$$X \in \mathbb{F}^{M\times K}, \quad W \in \mathbb{F}^{K\times N}, \quad W_{jk} \in \mathbb{F}^{s_j \times r_k}, \quad i = \tau(p).$$

Type-i projection

$$y_i = x_i W_{ii}, \qquad y_j = x_i W_{ij} + x_j W_{jj} \quad (j \neq i).$$

- Uses every diagonal block and the off-diagonal blocks in source row i.
- The full bank has KN learned scalars, covered jointly across types.
- $\tau(p)$ is a fixed position schedule; each projection is token-local.

Paper, Sec. 4, Eqs. (16)–(21), PDF pp. 4–5. M: token rows; K, N: feature widths.

Full matrix rank can coexist with restricted interactions

Teaching example, type 1:

$$(y_1, y_2) = (x_1, x_2) \underbrace{\begin{bmatrix} W_{11} & W_{12} \\ 0 & W_{22} \end{bmatrix}}_{T_1(W)}.$$

- Invertible diagonal blocks make $T_1(W)$ full rank.
- Yet x_2 cannot directly affect y_1.
- Another row type exposes W_{21}.

What associativity supplies

Compatible linear maps compose within the same family:

$$\Phi_{W_2} \circ \Phi_{W_1} = \Phi_{W_1 \star W_2}.$$

Equal parameter storage still allows a different function class.

Teaching representation derived from Eqs. (16)–(17). Paper, Secs. 2.1 and 4, PDF pp. 3–5.

The trained recursive law differs from direct four-group algebra

Law	Slots	Displayed products	Projection fraction
$\mathcal{Q}_2$	4	6	3/4
$\mathcal{Q}_4$	16	28	7/16
$\mathcal{Q}_2^{\otimes 2}$	16	36	9/16

- $\mathcal{Q}_4$ has exact rank 28; the recursive law has rank in $[28, 36]$.
- Recursion retains some interactions through two intermediate steps.
- The LM pilot uses $\mathcal{Q}_2^{\otimes 2}$: 43.75% fewer FFN MACs.

Paper, Sec. 3.2, PDF p. 4; App. C, pp. 12–13. Product counts refer to the full block table.

Quadratic arithmetic requires a growing number of groups

Equal-width rectangular projections:

$$C_{\mathrm{proj}} = \rho_q MKN, \qquad \rho_q = \frac{2q-1}{q^2}.$$

For square operands, $n = qb$ with physical block size b:

$$C(n,q) = (2q^2 - q)(n/q)^3.$$

Fixed q

$C(n,q) = \Theta(n^3)$

A constant-factor reduction.

Fixed b, $q = n/b$

$C = 2bn^2 - b^2 n = \Theta(n^2)$

The interaction law changes with n.

Paper, Sec. 5.1, Eqs. (22)–(24), PDF p. 5; App. H.2, p. 22. Counts are MACs; storage remains n^2.

GPU efficiency determines where subdivision should stop

Type-specific GEMMs have approximate dimensions

$$(M/q) \times (K/q) \times (N/q).$$

- Small blocks reduce arithmetic but may lose utilization or waste padded tiles.
- Fused endpoint accumulation and a two-stage schedule offer different reuse.
- One-token decoding uses a dedicated vector-reduction kernel.

Compute-dominated crossover model

$$\frac{\rho_q}{\eta_q} + \delta_q < 1, \qquad \eta_q = \frac{P_a(q)}{P_d}.$$

η_q: relative effective compute rate; δ_q: overhead / dense time.

Paper, Secs. 5.2–5.3, Eqs. (25)–(26), PDF pp. 5–6. A deployed architecture fixes q; kernel tiles and schedules may adapt.

The same row action supports training and cached decoding

Parallel training

Group token rows by $\tau(p)$.

Evaluate permitted rectangular GEMMs.

dX and dW reuse the same support:

$$C_{\text{train}} = 3C_{\text{fwd}}.$$

Autoregressive decoding

Compute only the new token row.

Preserve the absolute-position type.

Standard triangular attention and full KV cache remain valid.

Causality follows from token-local projections and the temporal mask.

Paper, Sec. 6, Algorithm 1, PDF pp. 6–7; App. B, p. 12; App. H.3, p. 22. The three-pass count covers matrix-product MACs.

The evidence comes from three separate settings

Setting	Object	What it measures
Theory	Direct $\mathcal{Q}_q$ family	Exact rank and arithmetic scaling
H200 kernels	Synthetic BF16 inputs; public layer dimensions	Projection, FFN, and gradient latency
Small LM	Recursive $\mathcal{Q}_2^{\otimes 2}$; 110.6M parameters	Generation throughput and task quality

- Kernel calibration is separated from alternating timing trials.
- No Qwen or DeepSeek model weights are loaded in the shape benchmarks.
- Large-q kernel results have no matching trained-model quality study.

Paper, Sec. 7, PDF pp. 7–8; Apps. E–F, pp. 15–16.

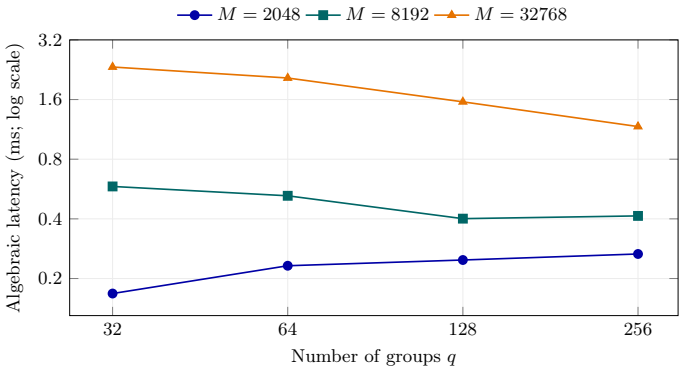
Pure projections accelerate across several large shapes

Shape	M	Selected q	Warm	Flushed
Qwen3-1.7B	2048	32	3.04×	2.84×
Qwen3-4B	2048	32	3.14×	2.94×
Qwen3-8B	2048	32	5.04×	4.52×
Qwen3-30B expert	2048	16	1.57×	1.30×
DeepSeek-V3 expert	2048	16	2.91×	2.47×
Qwen3-32B	32768	256	11.51×	11.70×

BF16, H200 NVL; speedup = dense kernel time / algebraic kernel time.
These timings isolate projections; MoE dispatch and communication are excluded.

Selected rows of paper Table 1, PDF p. 8. q minimizes warm latency in the tested grid; 256 is the largest tested value for Qwen3-32B.

The best group count depends on the number of token rows



$5120 \rightarrow 25600$

Qwen3-32B shape

Best tested q:

$M = 2048$: 32

$M = 8192$: 128

$M = 32768$: 256

More groups can increase latency.

Redrawn from paper Table 12, PDF p. 17. H200 NVL, BF16, warm medians; logarithmic latency axis.

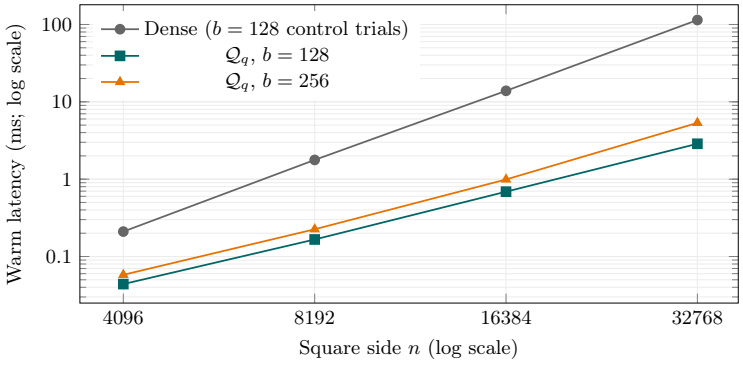
Forward gains can disappear in the gradient kernels

Shape	Full FFN forward	Up: fwd+dX+dW	Down: fwd+dX+dW
Qwen3-1.7B	2.55×	1.27×	1.76×
Qwen3-8B	4.17×	1.96×	2.56×
Qwen3-30B expert	1.74×	0.65×	0.57×
DeepSeek-V3 expert	3.74×	1.79×	1.35×

- All rows: $M = 2048$, $q = 32$, warm replay on H200 NVL.
- The small expert's down projection rises from 38.5 to 67.6 μs.
- Only forward schedules are calibrated; gradient tiles are fixed.

Selected rows of paper Tables 9–10, PDF p. 16; protocol in App. E, p. 15. Full FFN and individual projection training are different workloads.

Fixed leaf blocks show near-quadratic latency over four sizes



$q = n/b$ grows.

For $b = 128$:
endpoint slope ≈ 2.01.

At $n = 32768$:
$39.85\times$ warm speedup.

Synthetic products;
model quality is
untested here.

Redrawn from paper Table 13, PDF p. 18; discussion in App. F, p. 16. Both axes logarithmic. Ratios use the paper's reported values.

The LM pilot changes only the FFN multiplication law

Shared architecture

110,631,168 parameters
12 pre-norm decoder blocks
$D = 768$, $F = 1536$, 32 heads
Bias-free SwiGLU
Context: 1536 tokens
GPT-2 vocabulary: 50,257

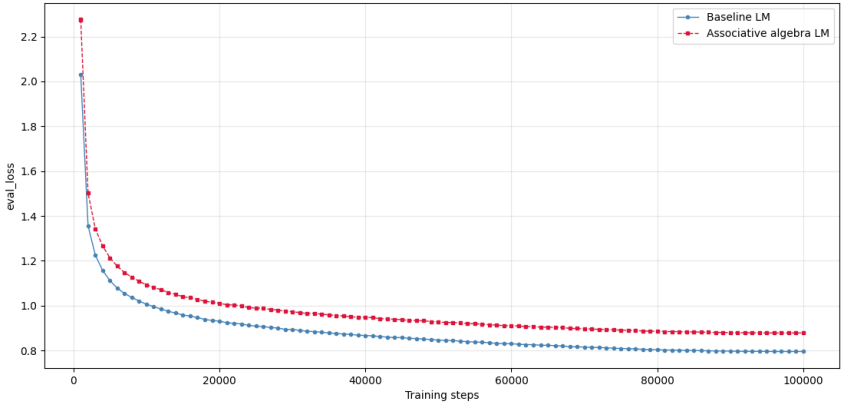
Shared training recipe

One run per model, from scratch
12.288B token slots each
100,000 steps on 2 H100s
AdamW, peak LR 3×10^{-4}
Same public instruction mixture
BF16 computation; FP32 parameters

Dense attention, embeddings, and LM head in both models.
Algebraic FFN: $\mathcal{Q}_2^{\otimes 2}$, projection MAC fraction 9/16.

Paper, Sec. 7.2, PDF pp. 7–8; Table 4 and App. D.1, p. 13.

Both models learn, while the algebraic loss stays higher



Both losses decrease; the algebraic curve remains higher.

Paper Fig. 1, PDF p. 13. Original author image; no exact endpoint loss values are tabulated.

Realized generation throughput improves by about 7%

Domain	Dense tok/s	Algebraic tok/s	Relative gain
Code	987.8	1054.5	6.8%
Math	991.4	1060.6	7.0%
QA	958.6	1033.1	7.8%
WMT ru–en	985.9	1047.1	6.2%

- 100 prompts/domain; three retained runs after one warm-up.
- Includes tokenization and sampling; both use CUDA-graph decoding.
- Response lengths differ: Code 195.5 vs. 182.4 tokens; QA 92.6 vs. 110.8.

Paper Table 2, PDF p. 8; Tables 5–7 and App. D.2, pp. 14–15. Tabulated means imply a 6.93% average gain. Fixed-length timings would isolate the layer effect.

All three downstream scores fall in the algebraic model

Metric (percent; higher is better)	Dense	Algebraic	Gap (pp)
GSM8K flexible EM	3.18 ± 0.48	2.20 ± 0.40	-0.98
MBPP pass@1	6.60 ± 1.11	3.80 ± 0.86	-2.80
IFEval loose accuracy	34.53	31.06	-3.47

$\pm$ denotes evaluation standard error, not variation across training runs.

A metric-label ambiguity in v1

Main Table 3 says *instance-level*; Appendix Table 8 says *instruction-level*. Values above follow the appendix.
The exact IFEval aggregation key remains unclear from the paper.

Paper Table 3, PDF p. 9; Table 8, p. 15. GSM8K: 5-shot; MBPP/IFEval: zero-shot. Gaps use rounded appendix percentages.

FFN arithmetic savings are diluted by the rest of the model

With dense attention and algebraic FFN projections:

$$C_{\mathrm{dec}} = \underbrace{2DH_q + 2DH_{kv}}_{\text{attention projections}} + \underbrace{\rho(g+1)DF}_{\text{FFN}} + \underbrace{2NH_q}_{\text{attention contractions}} .$$

Pilot: $H_q = H_{kv} = D$, $F = 2D$, $g = 2$, $\rho = 9/16$.

$$\frac{\text{algebraic block projection MACs}}{\text{dense block projection MACs}} = \frac{4D^2 + (9/16)6D^2}{10D^2} = \frac{59}{80} = 73.75\%.$$

- FFN MAC reduction: 43.75%; block projection reduction: 26.25%.
- Attention contractions, LM head, memory traffic, and sampling add cost.
- Parameter storage and KV-cache size remain unchanged.

Paper, Sec. 6.3, Eq. (29), PDF p. 7; Sec. 9, p. 9. N: cached positions; H_q, H_{kv}: head widths.

The contribution is a structured efficiency–quality trade-off

Established by the paper

A rank-certified interaction family, a causal row implementation, and measurable kernel gains. A recursive small LM trains successfully and generates faster, with lower task scores.

What changes the practical verdict

Large-q quality, full-model training time, and equal-quality cost have not been established. The two LMs have one training run each.

Discussion proposal: compare direct and recursive four-group laws under one recipe, then measure fixed-length latency and cost at matched quality.

Paper, Sec. 9–10, PDF p. 9. Discussion proposal is reader analysis.

Paper and navigation

Ilya Koziev and Ivan Oseledets

Change the Product, Keep the Parameters:
Associative Algebra Layers for Transformers

arXiv:2609.32814v1, 26 September 2026
https://arxiv.org/abs/2609.32814v1

- Core construction and costs: Secs. 2–6; proofs in Apps. A–C and H.
- Main evidence: Sec. 7; detailed results in Apps. D–F.
- Proposed attention extension: App. G; data description: App. I.

All empirical data in this deck come from this version. PDF locators count from page 1. The next seven slides are optional technical backup.

Backup: associativity follows from commuting endpoint actions

Write an element as diagonal slots d and edge slots n:

$$(d, n) \star (d', n') = (dd',\ dn' + nd').$$

The diagonal component of either bracketing is $dd'd''$; the edge component is

$$dd'n'' + dn'd'' + nd'd''.$$

- Left diagonal multiplication acts at the source vertex.
- Right diagonal multiplication acts at the target vertex.
- These actions commute; the identity is $(\mathbf{1}, 0)$.

Paper, App. A, Eq. (30), PDF p. 12. For the graph proof and exact rank, see App. H.1, pp. 21–22.

Backup: backward uses the same interaction support

For a token of type i, with output gradient $\bar{y}_j$:

$$\bar{x}_i \mathrel{+}= \bar{y}_i W_{ii}^\top + \sum_{j \neq i} \bar{y}_j W_{ij}^\top, \qquad \bar{x}_j \mathrel{+}= \bar{y}_j W_{jj}^\top \quad (j \neq i),$$

$$\bar{W}_{ii} \mathrel{+}= x_i^\top \bar{y}_i, \qquad \bar{W}_{ij} \mathrel{+}= x_i^\top \bar{y}_j \quad (j \neq i),$$

$$\bar{W}_{jj} \mathrel{+}= x_j^\top \bar{y}_j \quad (j \neq i).$$

Unequal-width exact count:

$$C_i = \sum_j s_j r_j + s_i(N - r_i), \qquad C_{\text{train}} = 3 \sum_i M_i C_i.$$

Paper, App. B, Eqs. (31)–(35), PDF p. 12; Sec. 4, Eqs. (18)–(20), p. 4.

Backup: recursive endpoints and the actual position schedule

Types and group indices are L-bit strings:

$$y_\delta = \sum_{\beta:\ \beta_k \in \{\alpha_k, \delta_k\}} x_\beta W_{\beta\delta}.$$

- Hamming distance $h(\alpha, \delta)$ gives 2^h intermediate groups.
- At $L = 2$: full-table count $4 \cdot 1 + 8 \cdot 2 + 4 \cdot 4 = 36$.
- Equal-width per-token fraction: $(3/4)^L$.

Pilot schedule, zero-based positions

$$\tau(p) = \left\lfloor \frac{p \bmod 768}{192} \right\rfloor.$$

Consecutive 192-token intervals share a type; kernel microbenchmarks use cyclic types.

Paper, App. C, Eqs. (36)–(37), PDF pp. 12–13; App. D.5, p. 15.

Backup: structured attention scores require support-aware scaling

This is a proposed extension; the LM pilot keeps attention dense.

$$F(i,j) = \{i,j\}, \quad d_h = qm, \quad s_{pu} = \frac{\sum_{a \in F(i,j)} \langle Q_p^{(a)}, K_u^{(a)} \rangle}{\sqrt{|F(i,j)|m}}.$$

- $|F(i,j)| = 1$ for matching types; otherwise 2.
- Scaling assumes selected coordinates are initially independent, centered, and unit variance.
- Standard causal softmax and dense PV follow the score calculation.

Paper, App. G.1, Eqs. (38)–(44), PDF pp. 18–19. i, j are the query/key position types; m is the width of one feature group.

Backup: structured scores still require the full KV cache

A future query type may access a different group of the same stored key.

$$K^{\text{cache}}, V^{\text{cache}} \in \mathbb{R}^{N_p \times n_{kv} \times d_h}, \qquad \text{storage} = 2N_p H_{kv}.$$

If the cache contains n_i keys matching query type i:

$$C_{QK}^{(i)} = \frac{H_q}{q}(2N_p - n_i), \qquad C_{PV} = N_p H_q.$$

Law	Score fraction	Forward ceiling	Training ceiling
$\mathcal{Q}_4$	$7/16$	$1.391\times$	$1.474\times$
$\mathcal{Q}_2^{\otimes 2}$	$9/16$	$1.280\times$	$1.333\times$

Paper, App. G.2, Eqs. (52)–(61), PDF pp. 20–21; Table 14, p. 20. Arithmetic ceilings for attention contractions; dense PV retained, with score recomputation in training.

Backup: generation and training use distinct protocols

Training data

Public instruction mixture.

Exact query–response deduplication.

Benchmark near-duplicate checks via 3-character shingles; Jaccard > 0.80 removed.

Extra MBPP solution checks for code responses.

Evaluation

Throughput: temperature 0.8, top-p 0.95, top-k 50; up to 256 tokens.

Quality: harness 0.4.9.1, batch 1, BF16; up to 512 tokens.

Timing CIs use 3 retained runs; task SEs use evaluation examples.

Paper, Apps. D.2–D.4, PDF p. 14; App. I, pp. 23–24. The dataset manifest counts imported records, not consumed token proportions.

Backup: full-FFN gains vary with cache reuse

Fixed $q = 32$, complete SwiGLU forward; speedups are warm / flushed.

Shape	$M = 1$	$M = 2048$
Qwen3-1.7B	$3.45\times$ / $2.46\times$	$2.55\times$ / $2.45\times$
Qwen3-4B	$4.74\times$ / $3.21\times$	$2.78\times$ / $2.68\times$
Qwen3-8B	$6.64\times$ / $4.12\times$	$4.17\times$ / $3.92\times$
Qwen3-30B-A3B expert	$1.71\times$ / $1.42\times$	$1.74\times$ / $1.54\times$
DeepSeek-V3 expert	$3.94\times$ / $2.75\times$	$3.74\times$ / $3.40\times$

- Both methods merge gate/up and use the same activation kernel.
- At $M = 1$, warm graph replay repeatedly uses one row type.
- These are isolated FFNs on synthetic BF16 tensors, on H200 NVL.

Paper, Table 9, PDF p. 16; measurement protocol in App. E, p. 15.